Title
Big Data, Little Data, or noData? Sustaining Access to Scholarship

Author
Borgman, Christine L.

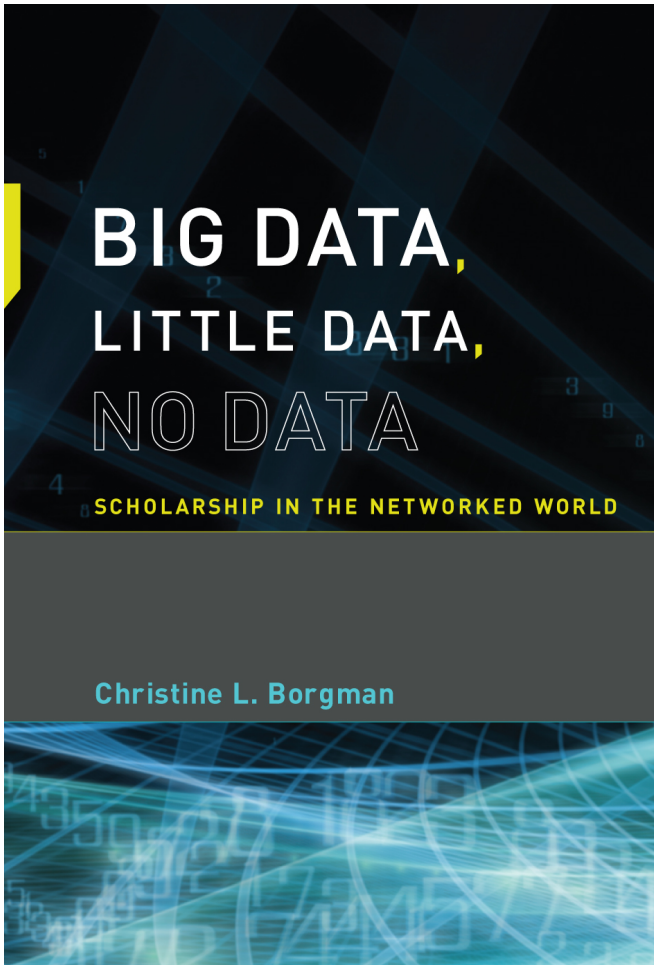
Publication Date
2016-12-06

Big Data, Little Data, or No Data? Sustaining Access to Scholarship

Christine L. Borgman

Distinguished Professor and Presidential Chair in Information Studies

University of California, Los Angeles



Distinguished Lecture

Academia Sinica, Taipei

6 December 2016

PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD

Vol I.
For *Anno* 1665, and 1666.

In the *SAVOY*,
Printed by *T. N.* for *John Martyn* at the Bell, a little without *Temple-Bar*, and *James Allestry* in *Duck-Lane*,
Printers to the *Royal Society*.

Theme issue 'Celebrating 350 years of Philosophical Transactions: life sciences papers' compiled and edited by Linda Partridge
19 April 2015; volume 370, issue 1666

Open access policies

- Research Councils of the UK
- European Union
- Australian Research Council
- U.S. Federal research policy
- Taiwan, China, India…
- Individual countries, funding agencies

Supported by wellcome trust

Australian Government
National Health and Medical Research Council

National Science Foundation
WHERE DISCOVERIES BEGIN

NIH National Institutes of Health
Turning Discovery Into Health

Policy RECommendations for Open Access to Research Data in Europe

RECODE

eScholarship
University of California

eScholarship provides Open-Access scholarly publishing services to the University of California and delivers a dynamic research platform to scholars worldwide. Powered by the California Digital Library.

My Account | My Saved Items (0)

Search publications... GO

About eScholarship | Deposit/Publish | Browse Content | Help

UC OPEN ACCESS POLICY

Learn more ▶

DEPOSIT | WAIVER | FAQ

Popular Research This Month

1. Ebola Virus Disease: Essential Public He...
2. An Introduction To Green Marketing
3. How Languages are Learned
4. A Spectral Analysis of World GDP Dynamic...
5. Making Data Count: A Data Metrics Pilot ...

» SEE MORE

Manage Your Publications

Books
Publish online.
Print on demand.

Journals
Peer review and MS management made simple.

Working Papers
Share your research before it gets stale (or scooped).

Previously Published Works
Open your publications to the world.

Conferences
Plan your meeting.
Publish the results.

eScholarship by the Numbers

Views Since 2002:	23,216,720
Publications:	75,712
Research Units:	338
Journals:	70


Data

Research Data Sharing without barriers

Precondition:

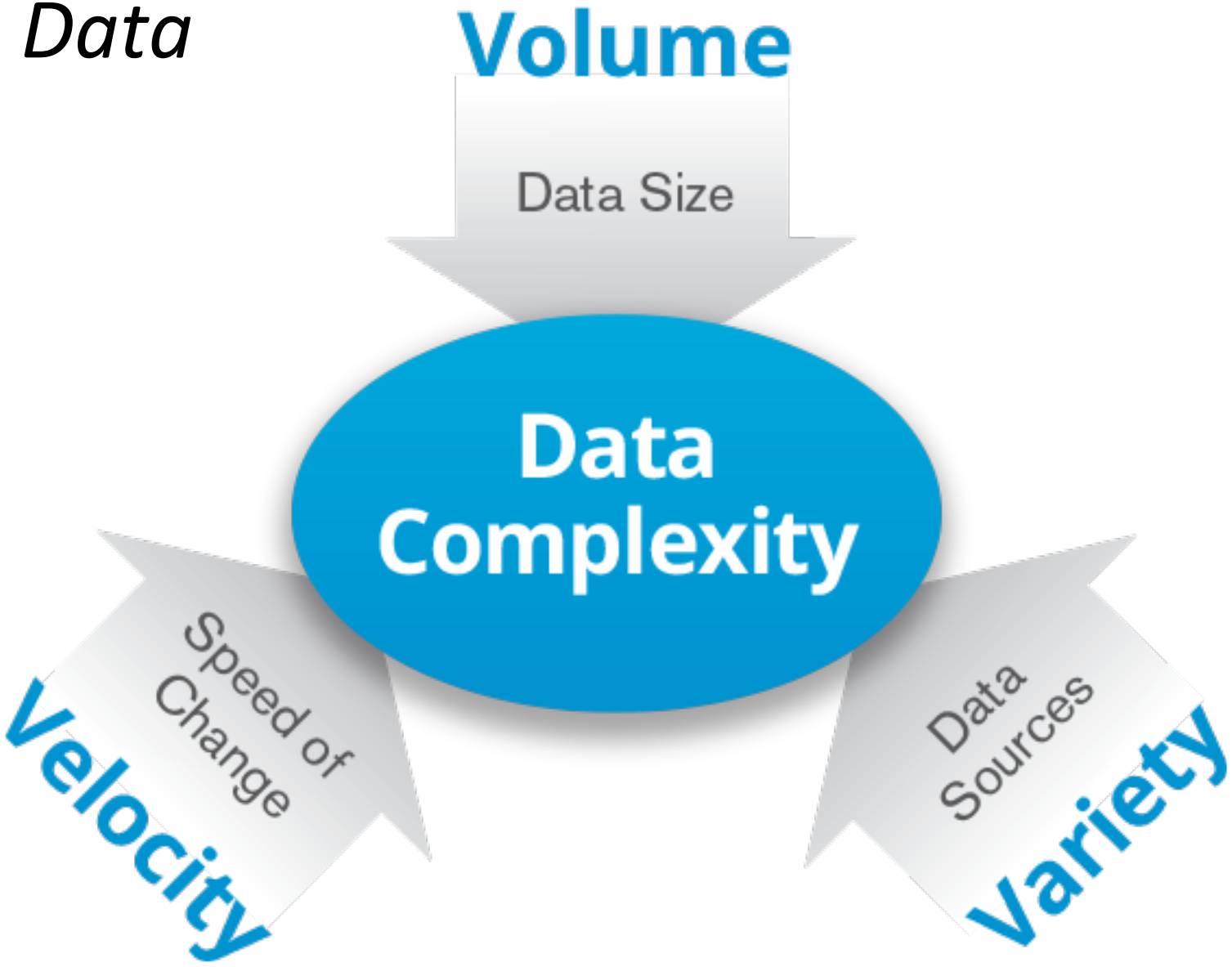
Researchers share data

Lack of incentives to share data

- Rewards for publication
- Effort to document data
- Competition, priority
- Control, ownership

http://www.buildingsrus.co.uk/.../ target1.htm

Big Data



http://www.datameer.com/product/hadoop.html

WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Illustration: Marian Bantjes

How Data Failed Us in Calling an Election

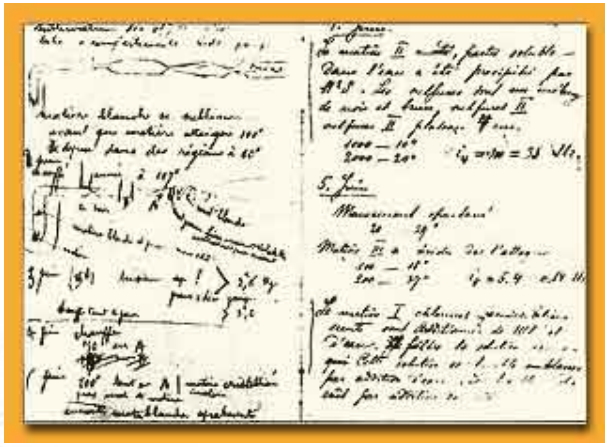
By **STEVE LOHR** and **NATASHA SINGER** NOV. 10, 2016

Michigan Democratic Party members in Flint studying precinct results on Tuesday. Virtually all major vote forecasters put Hillary Clinton's chances of winning in the range of 70 to 99 percent.

Brittany Greeson for The New York Times

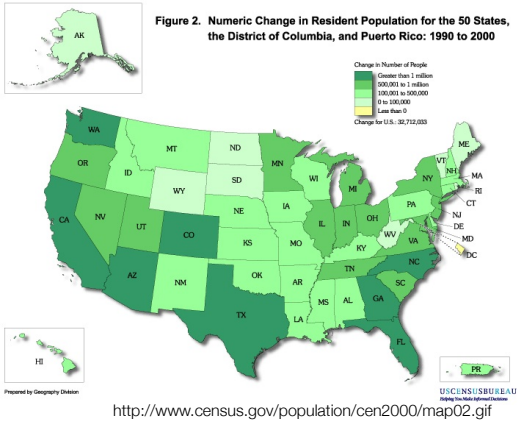
What are data?

hudsonalpha.org

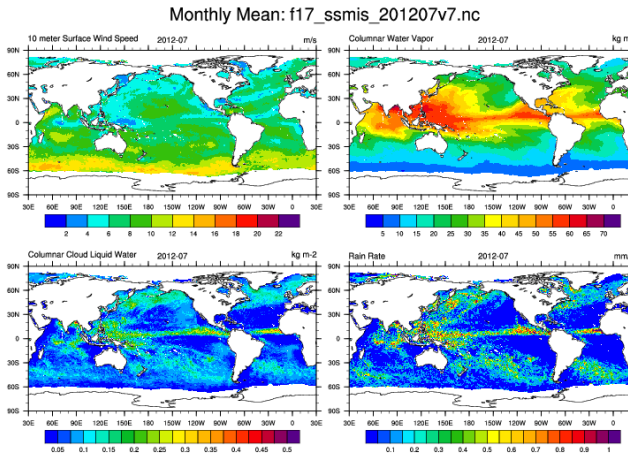


Marie Curie's notebook aip.org

Pisa Griffin



http://www.census.gov/population/cen2000/map02.gif



ncl.ucar.edu

Date:1/2.07.75 Place:Sakaltutan
Zafor
He will grow old in his present house; new house is for sons - 5 sons. Not sure they want to live in village. He will only build another if they want him to. eS came from Germany and did the plastering. He arranged the carpentry in Kayseri. Çok para gitti. {much money went} Has a tractor.

Date:July1980 Place:Sakaltutan
Zafor:
Household now Zafor and wife; Nazif Unal and wife and youngest son, still a boy. They run two dolmuß; one with a driver from Süleymanli. Goes in and out once a day. He gets 8,000 a month. Zafor then said, keskin de©il. { not sharp - i.e.? not profitable} I said he did very well on 8,000 TL with only two journeys a day. Nazif Unal has "bought" a Durak {dolmuß stop} from Belediye and works all day in Kayseri.

http://onlineqda.hud.ac.uk/Intro_QDA/Examples_of_Qualitative_Data.php

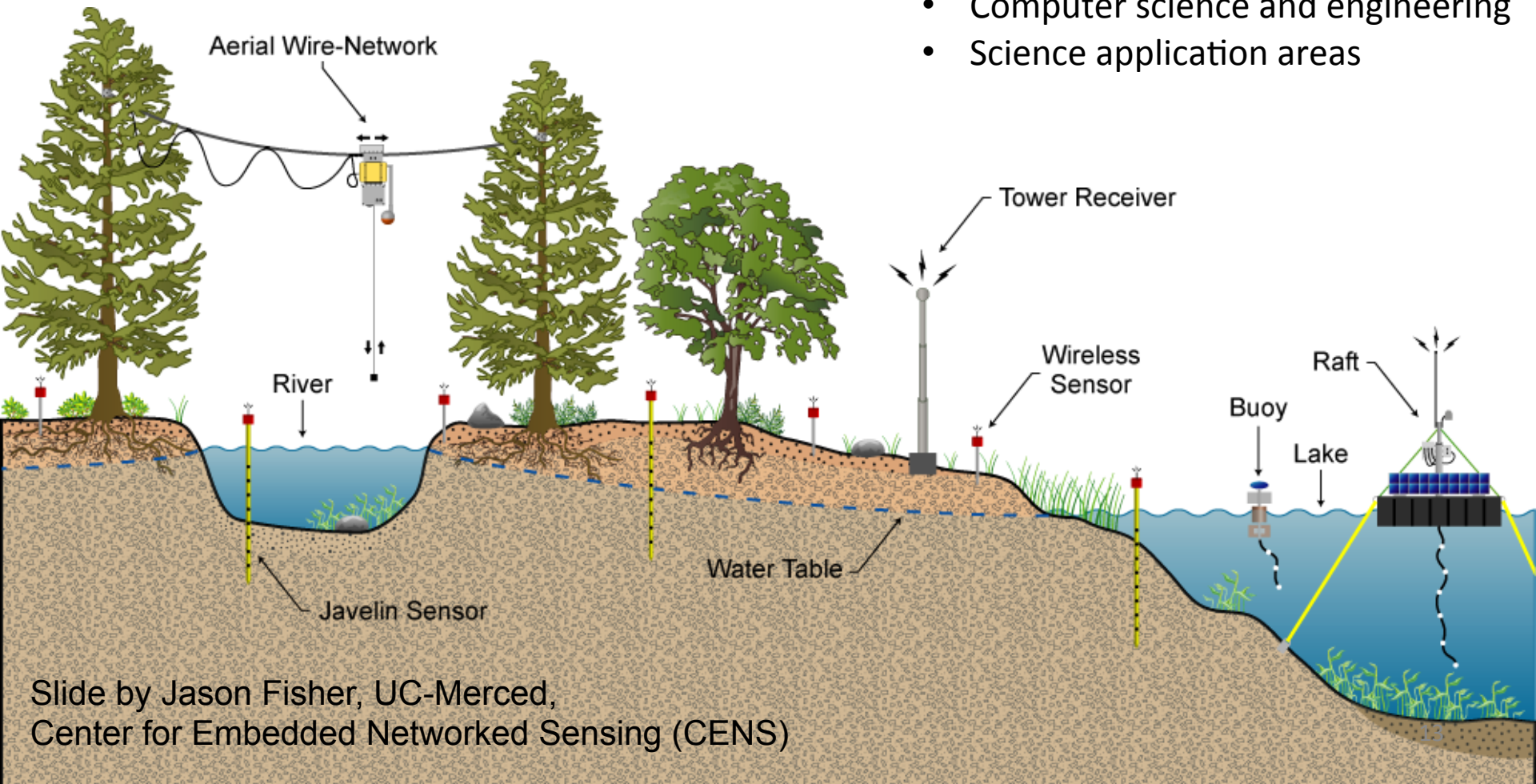
Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship.

C.L. Borgman (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press

http://www.genome.gov/dmd/img.cfm?node=Photos/Graphics&id=85327

Center for Embedded Networked Sensing

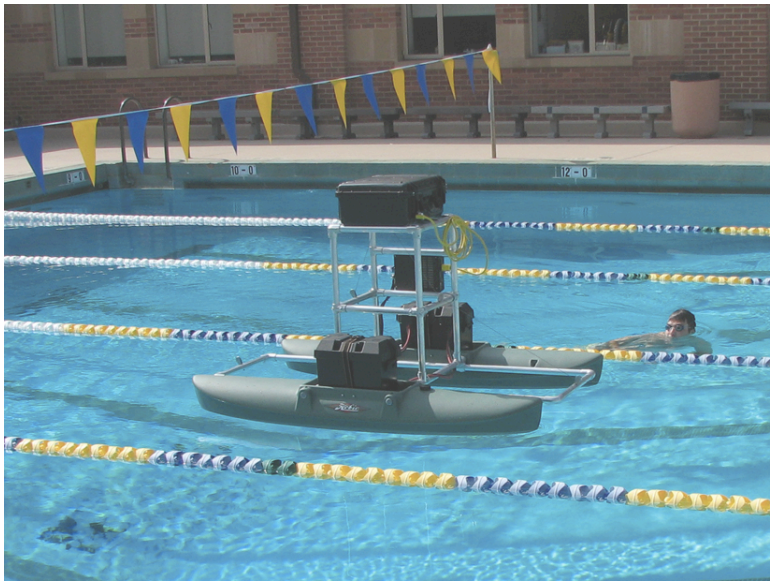
- NSF Science & Tech Ctr, 2002-2012
- 5 universities, plus partners
- 300 members
- Computer science and engineering
- Science application areas



Slide by Jason Fisher, UC-Merced,
Center for Embedded Networked Sensing (CENS)

Science <–> Data

Engineering researcher:
"Temperature is temperature."



CENS Robotics team

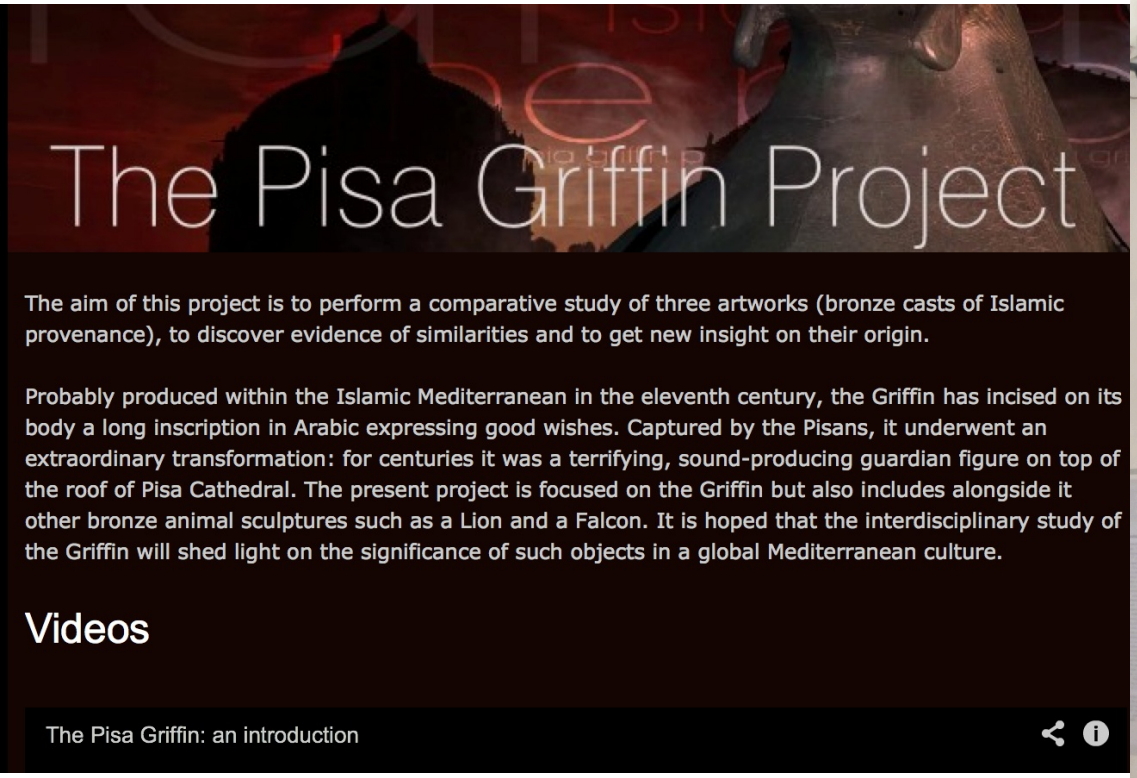
Biologist*:* ***"There are hundreds of ways to measure temperature.*** *'The temperature is 98' is low-value compared to, 'the temperature of the surface, measured by the infrared thermopile, model number XYZ, is 98.' That means it is measuring a proxy for a temperature, rather than being in contact with a probe, and it is measuring from a distance. The accuracy is plus or minus .05 of a degree. I [also] want to know that it was taken outside versus inside a controlled environment, how long it had been in place, and the last time it was calibrated, which might tell me whether it has drifted.."*

Center for Dark Energy Biosphere Investigations

International Ocean Discovery Program
Iodp.tamu.org

- NSF Science & Tech Ctr, 2010-2020
- 20 universities, plus partners (35 institutions)
- 90 scientists
- Biological sciences
- Physical sciences

Repository for seafloor cores. Photo: Peter Darch



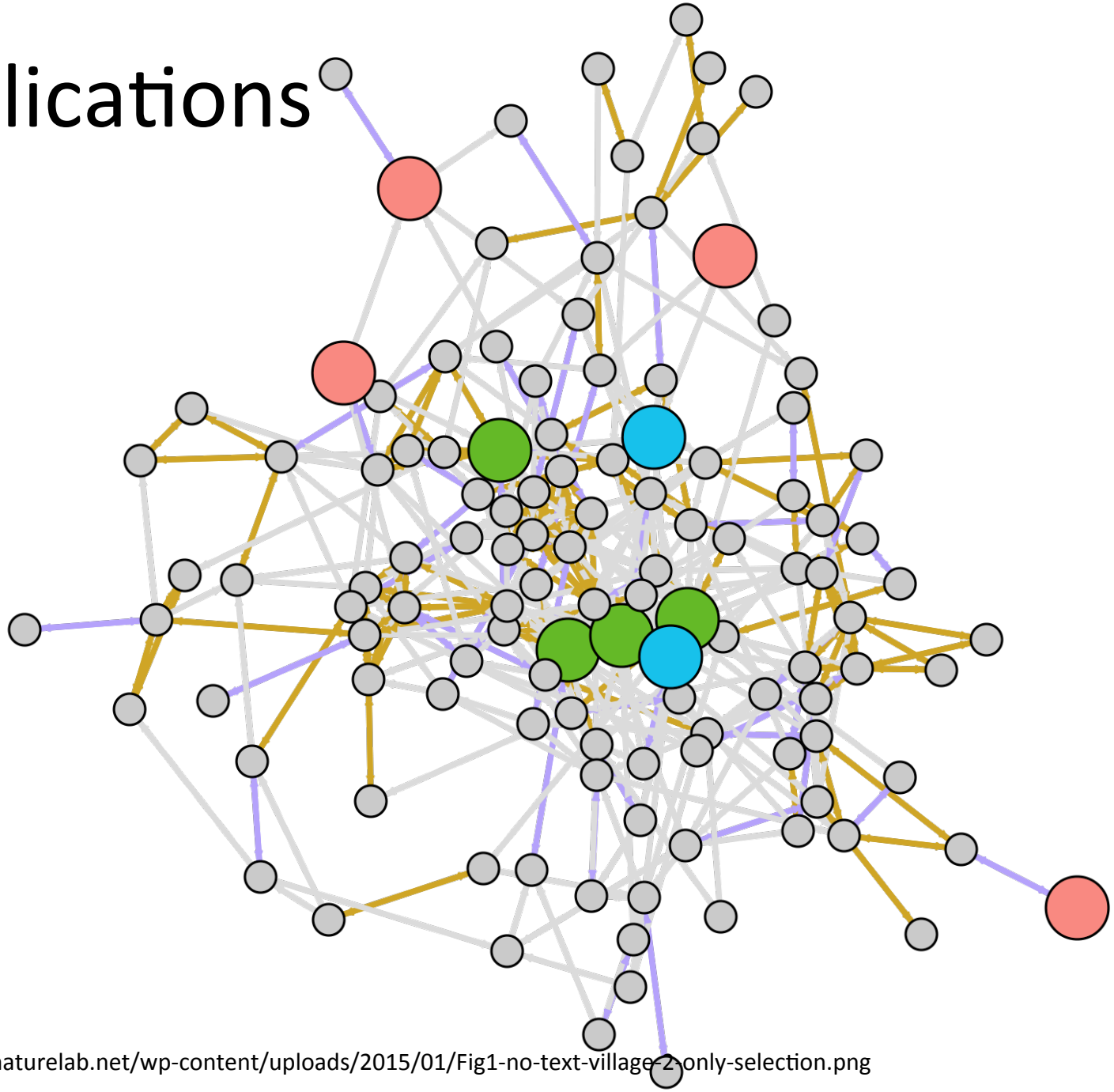
http://vcg.isti.cnr.it/griffin/

Arte islamica, ippogrifo, XI sec 03, own work

Publications

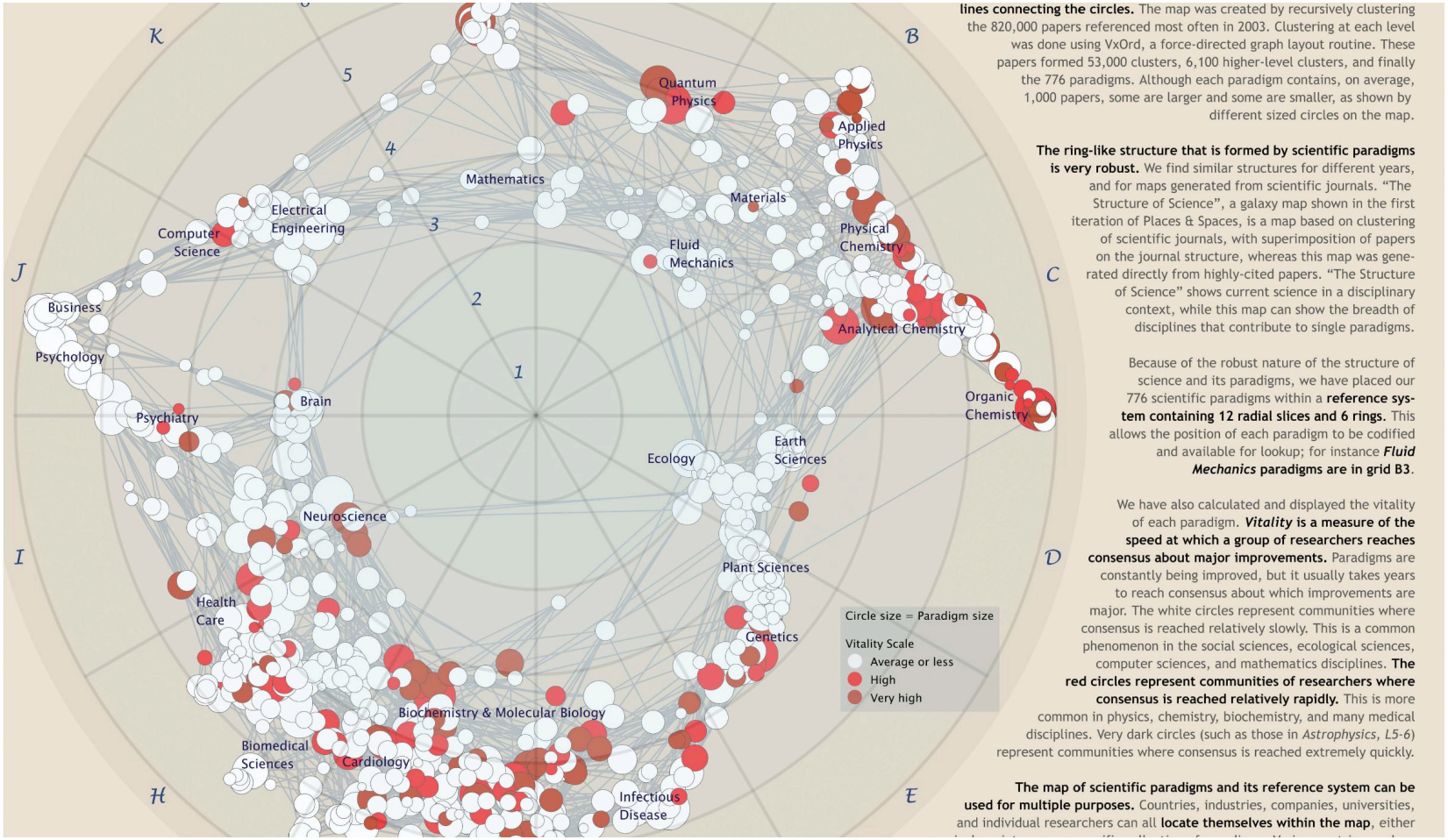
http://www.cse.psu.edu/hpcl/images/publications.jpg

Publications



http://humannaturelab.net/wp-content/uploads/2015/01/Fig1-no-text-village-2-only-selection.png

Mapping Scholarship



lines connecting the circles. The map was created by recursively clustering the 820,000 papers referenced most often in 2003. Clustering at each level was done using VxOrd, a force-directed graph layout routine. These papers formed 53,000 clusters, 6,100 higher-level clusters, and finally the 776 paradigms. Although each paradigm contains, on average, 1,000 papers, some are larger and some are smaller, as shown by different sized circles on the map.

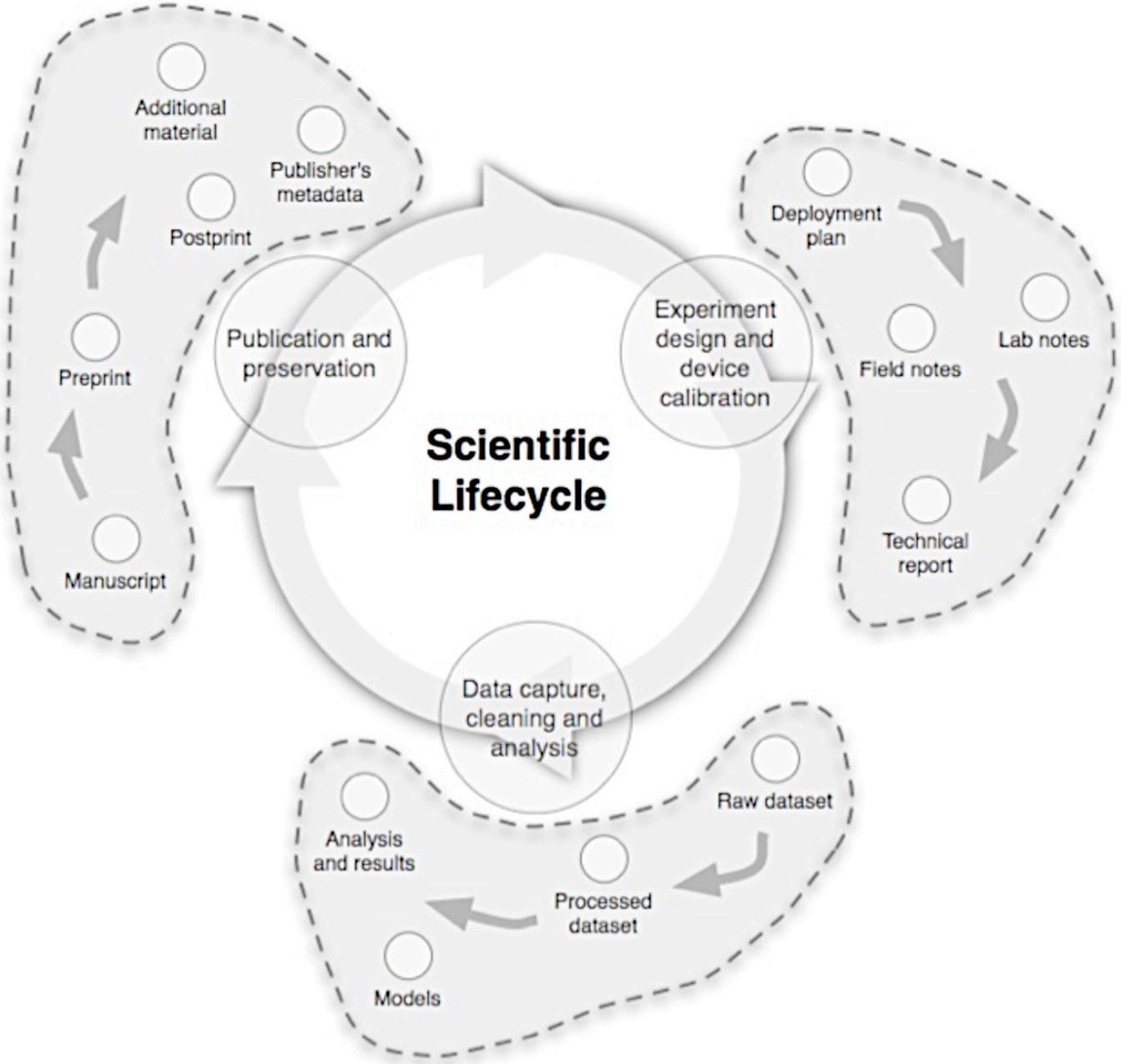
The ring-like structure that is formed by scientific paradigms is very robust. We find similar structures for different years, and for maps generated from scientific journals. "The Structure of Science", a galaxy map shown in the first iteration of Places & Spaces, is a map based on clustering of scientific journals, with superimposition of papers on the journal structure, whereas this map was generated directly from highly-cited papers. "The Structure of Science" shows current science in a disciplinary context, while this map can show the breadth of disciplines that contribute to single paradigms.

Because of the robust nature of the structure of science and its paradigms, we have placed our 776 scientific paradigms within a **reference system containing 12 radial slices and 6 rings.** This allows the position of each paradigm to be codified and available for lookup; for instance ***Fluid Mechanics* paradigms are in grid B3**.

We have also calculated and displayed the vitality of each paradigm. ***Vitality* is a measure of the speed at which a group of researchers reaches consensus about major improvements.** Paradigms are constantly being improved, but it usually takes years to reach consensus about which improvements are major. The white circles represent communities where consensus is reached relatively slowly. This is a common phenomenon in the social sciences, ecological sciences, computer sciences, and mathematics disciplines. **The red circles represent communities of researchers where consensus is reached relatively rapidly.** This is more common in physics, chemistry, biochemistry, and many medical disciplines. Very dark circles (such as those in *Astrophysics*, *L5-6*) represent communities where consensus is reached extremely quickly.

The map of scientific paradigms and its reference system can be used for multiple purposes. Countries, industries, companies, universities, and individual researchers can all **locate themselves within the map**, either

Börner, K. (2010). *Atlas of Science: Visualizing What We Know*. Cambridge, Mass: The MIT Press.



Pepe, A., Mayernik, M. S., Borgman, C. L. & Van de Sompel, H. (2010). From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web. Journal of the American Society for Information Science and Technology, 61(3): 567–582.

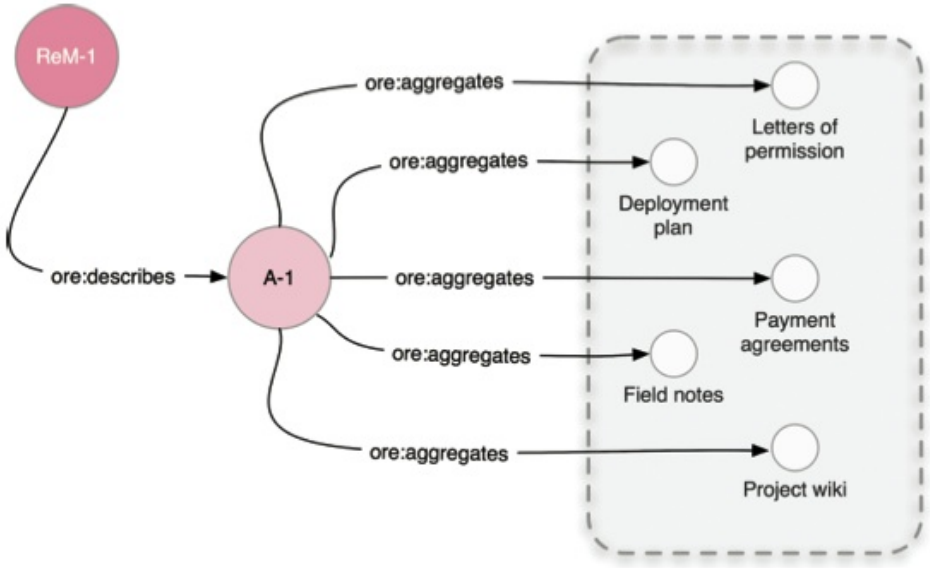


FIG. 4. ORE Aggregation representing the first stage of the scientific life cycle of a sensor network application in seismology (experiment and deployment planning).

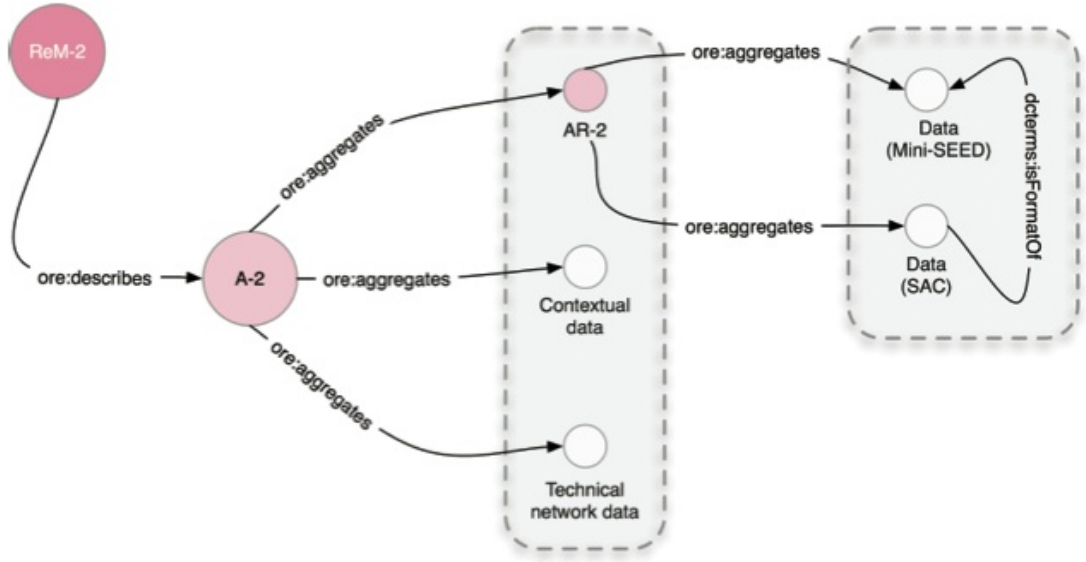


FIG. 5. ORE Aggregation representing the second stage of the scientific life cycle of a sensor network application in seismology (data collection).

Publications <–> Data: Role

Publications are arguments made by authors, and data are the evidence used to support the arguments.

C.L. Borgman (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press

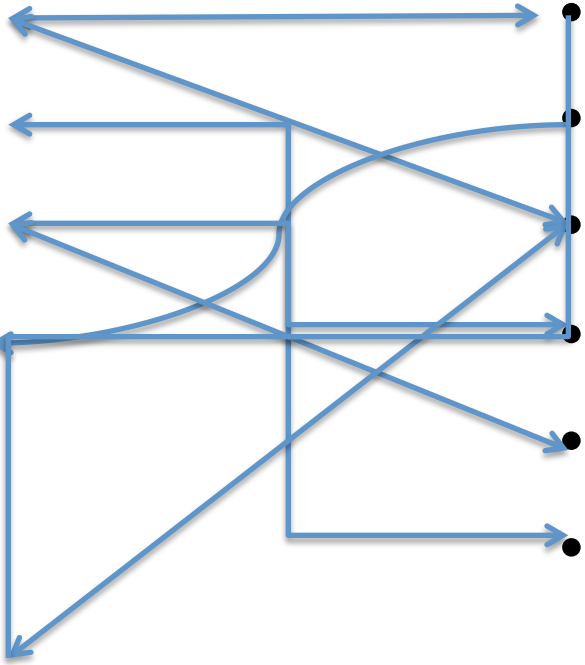
Publications <–> Data: Citations

"If publications are the stars and planets of the scientific universe, data are the 'dark matter' – influential but largely unobserved in our mapping process"*

*Micah Altman, CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013, p. 54

Publications <–> Data: Mapping

- Article 1
- Article 2
- Article 3
- Article 4
- Article n

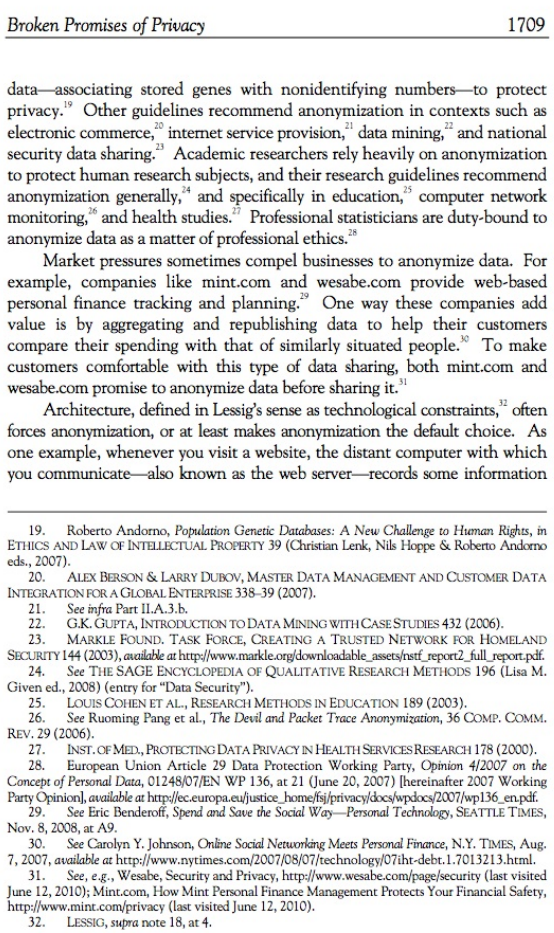


- Dataset time 1
- Dataset time 2
- Observation time 1
- Visualization time 3
- Community collection 1
- Repository 1

Data citation and analytics

- Credit
- Attribution
- Discovery

Bibliometrics, Scientometrics, Informetrics, Webometrics…



Broken Promises of Privacy 1709

data—associating stored genes with nonidentifying numbers—to protect privacy.[19] Other guidelines recommend anonymization in contexts such as electronic commerce,[20] internet service provision,[21] data mining,[22] and national security data sharing.[23] Academic researchers rely heavily on anonymization to protect human research subjects, and their research guidelines recommend anonymization generally,[24] and specifically in education,[25] computer network monitoring,[26] and health studies.[27] Professional statisticians are duty-bound to anonymize data as a matter of professional ethics.[28]

Market pressures sometimes compel businesses to anonymize data. For example, companies like mint.com and wesabe.com provide web-based personal finance tracking and planning.[29] One way these companies add value is by aggregating and republishing data to help their customers compare their spending with that of similarly situated people.[30] To make customers comfortable with this type of data sharing, both mint.com and wesabe.com promise to anonymize data before sharing it.[31]

Architecture, defined in Lessig's sense as technological constraints,[32] often forces anonymization, or at least makes anonymization the default choice. As one example, whenever you visit a website, the distant computer with which you communicate—also known as the web server—records some information

19. Roberto Andorno, *Population Genetic Databases: A New Challenge to Human Rights*, in ETHICS AND LAW OF INTELLECTUAL PROPERTY 39 (Christian Lenk, Nils Hoppe & Roberto Andorno eds., 2007).
20. ALEX BERSON & LARRY DUBOV, MASTER DATA MANAGEMENT AND CUSTOMER DATA INTEGRATION FOR A GLOBAL ENTERPRISE 338–39 (2007).
21. *See infra* Part II.A.3.b.
22. G.K. GUPTA, INTRODUCTION TO DATA MINING WITH CASE STUDIES 432 (2006).
23. MARKLE FOUND. TASK FORCE, CREATING A TRUSTED NETWORK FOR HOMELAND SECURITY 144 (2003), *available at* http://www.markle.org/downloadable_assets/nstf_report2_full_report.pdf.
24. *See* THE SAGE ENCYCLOPEDIA OF QUALITATIVE RESEARCH METHODS 196 (Lisa M. Given ed., 2008) (entry for "Data Security").
25. LOUIS COHEN ET AL., RESEARCH METHODS IN EDUCATION 189 (2003).
26. *See* Ruoming Pang et al., *The Devil and Packet Trace Anonymization*, 36 COMP. COMM. REV. 29 (2006).
27. INST. OF MED., PROTECTING DATA PRIVACY IN HEALTH SERVICES RESEARCH 178 (2000).
28. European Union Article 29 Data Protection Working Party, *Opinion 4/2007 on the Concept of Personal Data*, 01248/07/EN WP 136, at 21 (June 20, 2007) [hereinafter 2007 Working Party Opinion], *available at* http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/2007/wp136_en.pdf.
29. *See* Eric Benderoff, *Spend and Save the Social Way—Personal Technology*, SEATTLE TIMES, Nov. 8, 2008, at A9.
30. *See* Carolyn Y. Johnson, *Online Social Networking Meets Personal Finance*, N.Y. TIMES, Aug. 7, 2007, *available at* http://www.nytimes.com/2007/08/07/technology/07iht-debt.1.7013213.html.
31. *See, e.g.*, Wesabe, Security and Privacy, http://www.wesabe.com/page/security (last visited June 12, 2010); Mint.com, How Mint Personal Finance Management Protects Your Financial Safety, http://www.mint.com/privacy (last visited June 12, 2010).
32. LESSIG, *supra* note 18, at 4.

Ohm, P. (2010). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review*, *57*, 1701.

Aad, G., T. Abajyan, B. Abbott, J. Abdallah, S. Abdel Khalek, A. A. Abdelalim, O. Abdinov, et al. 2012. "Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC." *Physics Letters [Part B]* 716 (1):1–29. doi:10.1016/j.physletb.2012.08.020.

Abbate, Janet. 1999. *Inventing the Internet*. Cambridge, MA: MIT Press.

Accomazzi, Alberto. 2010. "Astronomy 3.0 Style." *Astronomical Society of the Pacific Conference Series* 433: 273–281.

Accomazzi, Alberto, and Rahul Dave. 2011. "Semantic Interlinking of Resources in the Virtual Observatory Era." *Astronomical Society of the Pacific Conference Series* 442: 415–424. doi: arXiv:1103.5958.

Acropolis Museum. 2013. "The Frieze." http://www.theacropolismuseum.gr/en/content/frieze-0.

Agosti, Maristella, and Nicola Ferro. 2007. "A Formal Model of Annotations of Digital Content." *ACM Transactions on Information Systems* 26 (1). doi:10.1145/1292591.1292594.

Agre, Philip E. 1994. "From High Tech to Human Tech: Empowerment, Measurement, and Social Studies of Computing." *Computer Supported Cooperative Work* 3 (2):167–195. doi:10.1007/BF00773446.

Ahn, Christopher P., Rachael Alexandroff, Carlos Allende Prieto, Scott F. Anderson, Timothy Anderton, Brett H. Andrews, Éric Aubourg, et al. 2012. "The Ninth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-III Baryon Oscillation Spectroscopic Survey." *Astrophysical Journal* 203:21. doi:10.1088/0067-0049/203/2/21.

Akyildiz, I. F., W. Su, Y. Sankarasubramaniam, and E. Cayirci. 2002. "Wireless Sensor Networks: A Survey." *Computer Networks* 38 (4):393–422. doi:10.1016/S1389-1286(01)00302-4.

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge MA: MIT Press.

Bibliometrics by Source

Searches for author: Christine Borgman, Christine L. Borgman, CL Borgman (excluding other C Borgman authors) on July 28, 2014 and November 26, 2016 for Google Scholar, Web of Science, Scopus
UCLA cancelled Scopus subscription by 2016

Source	Publications		Citations received		H-index	
	2014	2016	2014	2016	2014	2016
Google Scholar (Google)	380	443	7766	10714	39	45
Web of Science (Thomson-Reuters)	145	149*	1629	2124*	20	24
Scopus – July 2014 (Elsevier)	*77*		*1314*		*14 (after 1995)*	

*643 variants of cited publications were cited 3104 times

altmetrics

altmetrics: a manifesto

NO ONE CAN READ EVERYTHING. We rely on filters to make sense of the scholarly literature, but the narrow, traditional filters are being swamped. However, the growth of new, online scholarly tools allows us to make new filters; these altmetrics reflect the broad, rapid impact of scholarship in this burgeoning ecosystem. We call for more tools and research based on altmetrics.

As the volume of academic literature explodes, scholars rely on filters to select the most relevant and significant sources from the rest. Unfortunately, scholarship's three main filters for importance are failing:

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology

Jillian C. Wallis, Elizabeth Rolando, Christine L. Borgman

Published: July 23, 2013 • http://dx.doi.org/10.1371/journal.pone.0067332

Abstract

Research on practices to share and reuse data will inform the design of infrastructure to support data collection, management, and discovery in the long tail of science and technology. These are research domains in which data tend to be local in character, minimally structured, and minimally documented. We report on a ten-year study of the Center for Embedded Network Sensing (CENS), a National Science Foundation Science and Technology Center. We found that CENS researchers are willing to share their data, but few are asked to do so, and in only a few domain areas do their funders or journals require them to deposit data. Few repositories

Altmetrics data as of December 5, 2016: 49 cites per Scopus, 100 cites per Google Scholar; 50 cites per CrossRef

nature International weekly journal of science

NATURE | COMMENT

Bibliometrics: The Leiden Manifesto for research metrics

Diana Hicks, Paul Wouters, Ludo Waltman, Sarah de Rijcke & Ismael Rafols

22 April 2015

Use these ten principles to guide research evaluation, urge Diana Hicks, Paul Wouters and colleagues.

Subject terms: Careers · Research management · Publishing

San Francisco

DORA

Declaration on Research Assessment

Dora | Sign The Declaration | Inspiration and Good Practices | A Letter to Thompson Reuters

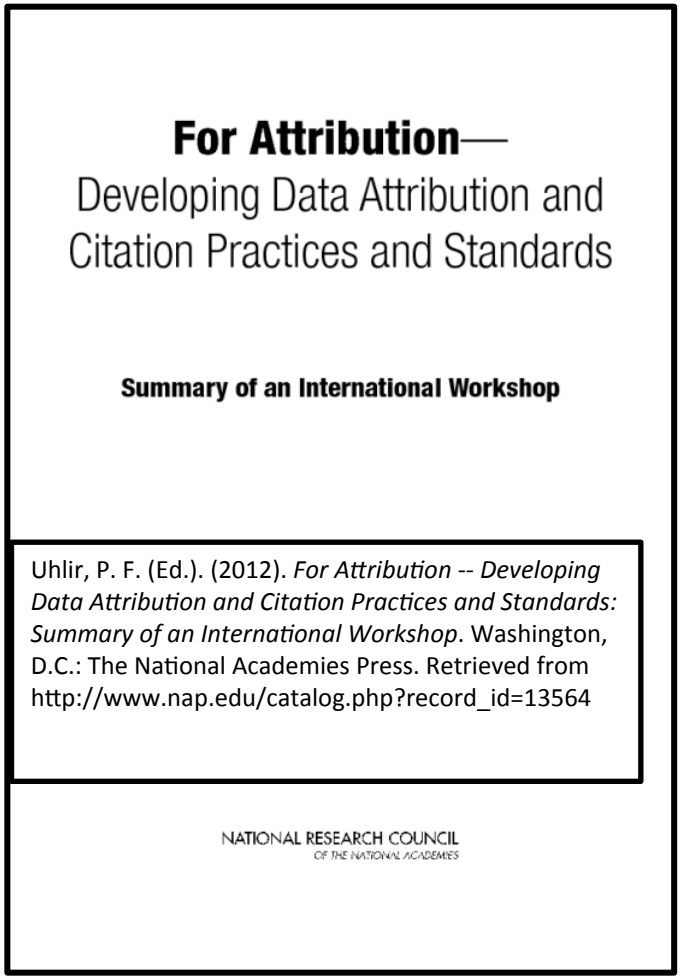
The San Francisco Declaration on Research Assessment (DORA), initiated by the American Society for Cell Biology (ASCB) together with a group of editors and publishers of scholarly journals, recognizes the need to improve the ways in which the outputs of scientific research are evaluated. The group met in December 2012 during the ASCB Annual Meeting in San Francisco and subsequently circulated a draft declaration among various stakeholders. DORA as it now stands has benefited from input by many of the original signers listed below. It is a worldwide initiative covering all scholarly disciplines. We encourage individuals and organizations who are concerned about the appropriate assessment of scientific research to sign DORA.

Publications <–> Data: Attribution

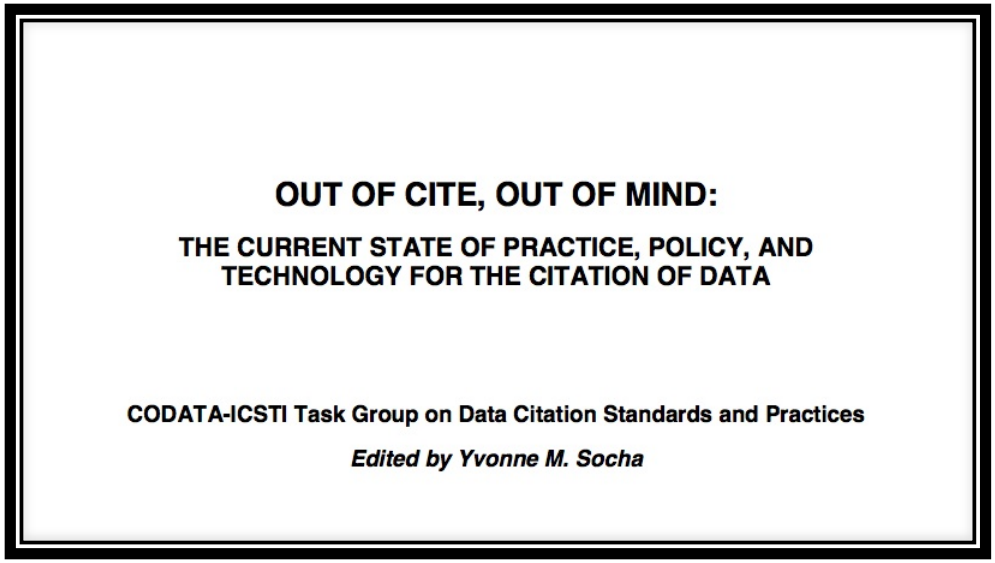
- Publications
  - Independent units
  - Authorship is negotiated
- Data
  - Compound objects
  - Ownership is rarely clear
  - Attribution
    - Long term responsibility: Investigators
    - Expertise for interpretation: Data collectors and analysts
- Software

http://www.genome.gov/dmd/img.cfm?node=Photos/Graphics&id=85327

Data Citation and Attribution



2012



Data Science Journal, Volume 12, 13 September 2013

CODATA-ICSTI Task Group on Data Citation and Attribution. Co-Chairs: Jan Brase, Sarah Callaghan, Christine Borgman

JOINT DECLARATION OF DATA CITATION PRINCIPLES - FINAL

When citing please use: Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 [**/datacitation**].

ENDORSEMENT LIST

PREAMBLE

Sound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record. In other words, data should be considered legitimate, citable products of research. Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse.

DC1
Data Citation Principles

In support of this assertion, and to encourage good practice, we offer a set of guiding principles for data within scholarly literature, another dataset, or any other research object.

These principles are the synthesis of work by a **number of groups**. As we move into the next phase, we welcome your participation and endorsement of these principles.

PRINCIPLES

The Data Citation Principles cover purpose, function and attributes of citations. These principles recognize the dual necessity of creating citation practices that are both human understandable and machine-actionable.

These citation principles are not comprehensive recommendations for data stewardship. And, as practices vary across communities and technologies will evolve over time, we do not include recommendations for specific implementations, but encourage communities to develop practices and tools that embody these principles.

The principles are grouped so as to facilitate understanding, rather than according to any perceived criteria of importance.

The principles are grouped so as to facilitate understanding, rather than according to any perceived criteria of importance.

1. ## Importance

   Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications[1].

2. ## Credit and Attribution

   Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data[2].

3. In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited[3].

4. ## Unique Identification

   A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community[4].

5. ## Access

   Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data[5].

6. ## Persistence

   Unique identifiers, and metadata describing the data, and its disposition, should persist -- even beyond the lifespan of the data they describe[6].

7. ## Specificity and Verifiability

   Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verfiying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited[7].

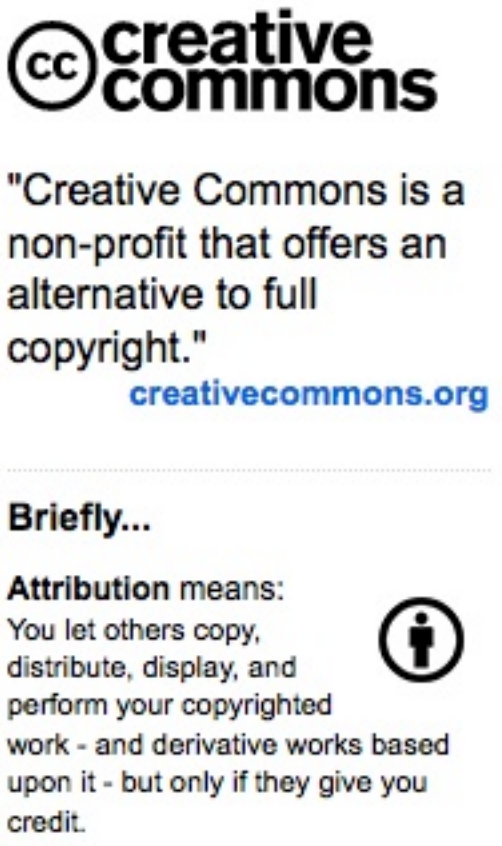
8. ## Interoperability and Flexibility

   Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities[8].

When citing this document please use: Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 [**https://www.force11.org/group/joint-declaration-data-citation-principles-final**].

Attribution of data

- Legal responsibility
  - Licensed data
  - Specific attribution required
- Scholarly credit: contributorship
  - "Author" of data
  - Contributor of data to this publication
  - Colleague who shared data
  - Software developer
  - Data collector
  - Instrument builder
  - Data curator
  - Data manager
  - Data scientist
  - Field site staff
  - Data calibration
  - Data analysis, visualization
  - Funding source
  - Data repository
  - Lab director
  - Principal investigator
  - University research office
  - Research subjects
  - Research workers, e.g., citizen science…



For Attribution -- Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop. Washington, D.C.: The National Academies Press. 2012

Discovery and Interpretation

- Identify the form and content
- Open
- Interpret
- Evaluate
- Compute upon
- Reuse
- Combine
- Describe
- Annotate…

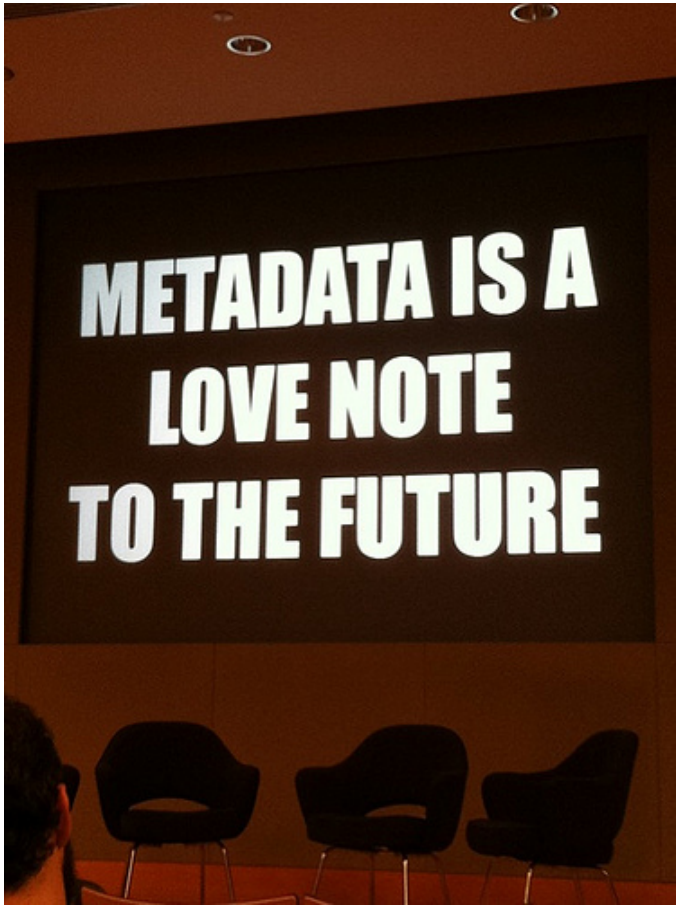
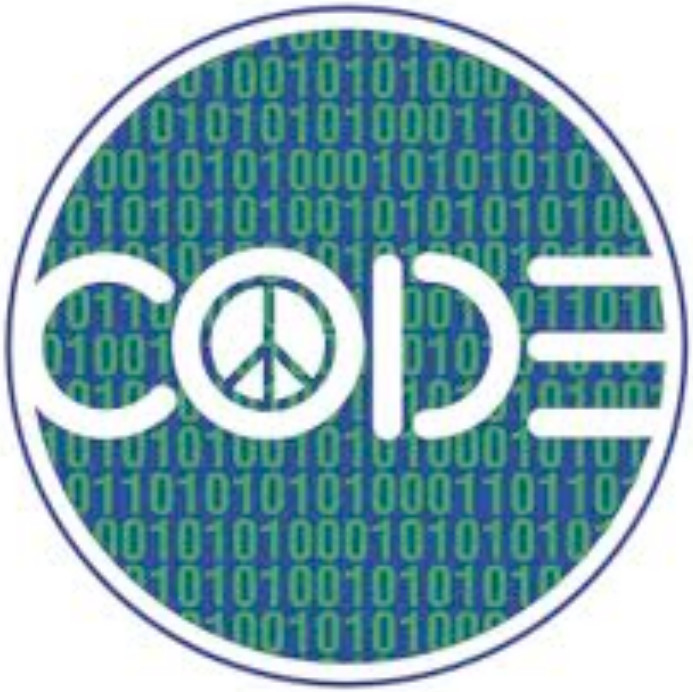


Photo by @kissane; presentation by Jason Scott (@textfiles)

Interpretation and replication

- Datasets
- Methods
  - Collection
  - Cleaning
  - Analysis
  - Codebook
- Publications
- Software and code
- Instrumentation



http://peacetour.org/sites/default/files/code4peace-logo2-v3-color-sm.jpg

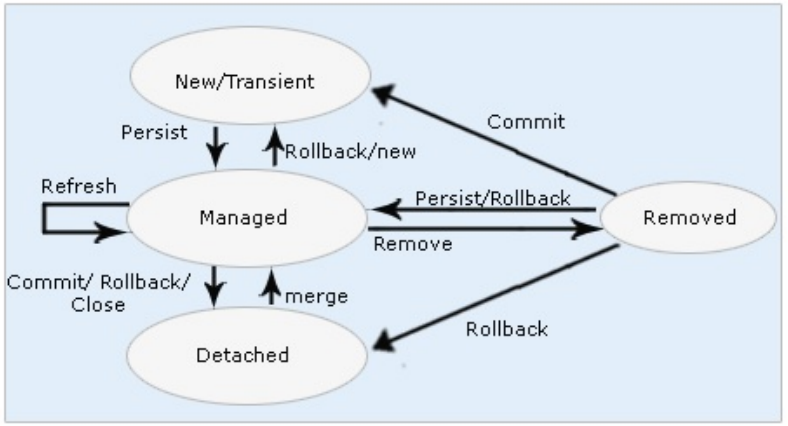
Intellectual property

- What can I do with this object?
- What rights are associated?
  - Reuse
  - Reproduce
  - Attribute
- Who owns the rights?
- How open are data?
  - Open data
  - Open bibliography

http://pzwart.wdka.hro.nl/mdr/research/lliang/mdr/mdr_images/opencontent.jpg/

Persistence and Identity

- Persistence
  - Perishable
  - Long-lived
  - Permanent
- Identity
  - Identifiers
    - DOI, Handles
    - URI, PURL…
  - Naming and namespaces
    - Authors/creators: ORCID, VIAF…
    - Generic/specific: registry number…
  - Description
    - Self-describing
    - Metadata augmentation

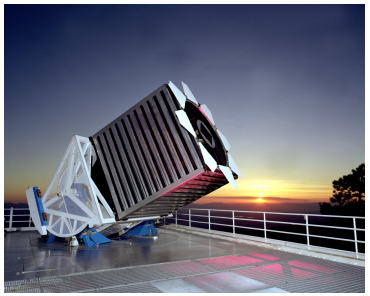


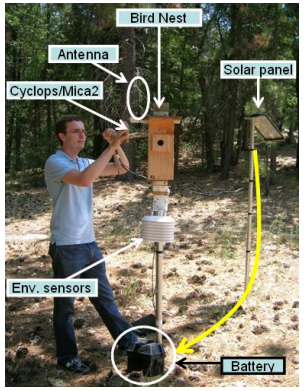
Persistence Content

http://web-interview-questions.blogspot.com/2010_06_21_archive.html

Some ways to release data

- Centralized data production
  - Top down investments in data
  - Common data archive
- Decentralized data production
  - Bottom up investments in data
  - Pool domain resources later
- Domain-independent aggregators
  - University repositories
  - Dataverse, Figshare, Slideshare, …
- Post on lab / personal websites
- Share privately upon request





MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

https://github.com/okulbilisim/awesome-datascience

Data Curation and Stewardship

- Services and tools
- Data management planning
- Selection and appraisal
- Metadata, provenance
- Migration
- Economics
- Infrastructure

http://www.librarygirl.net/2013/08/putting-your-best-foot-forward-tl.html

Reuse across place and time

- Reuse by investigator
- Reuse by collaborators
- Reuse by colleagues
- Reuse by unaffiliated others
- Reuse at later times
  - Months
  - Years
  - Decades
  - Centuries

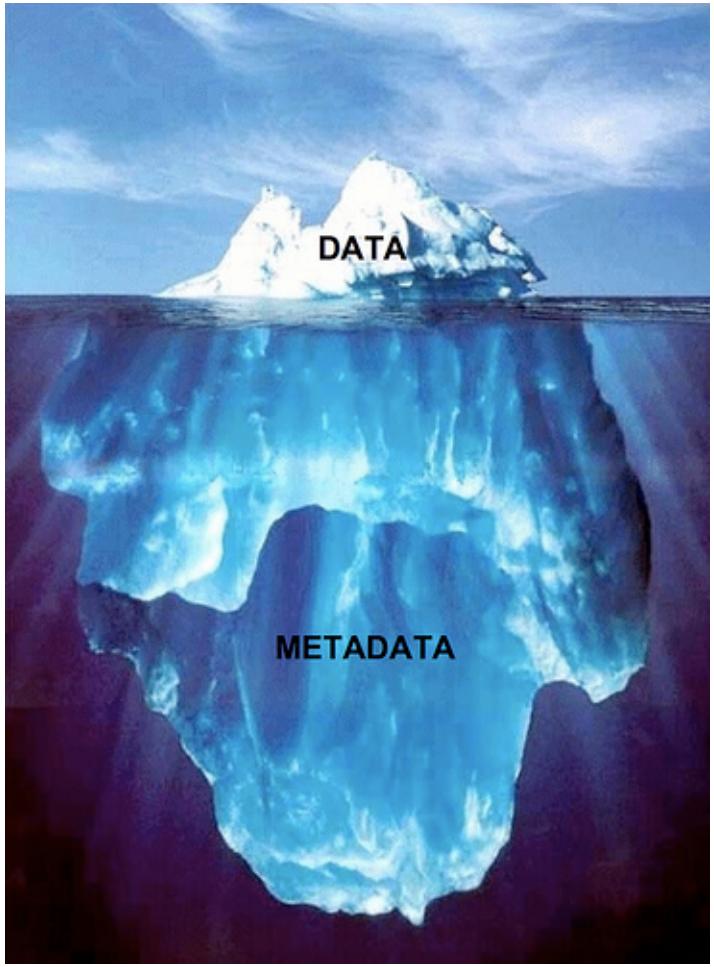
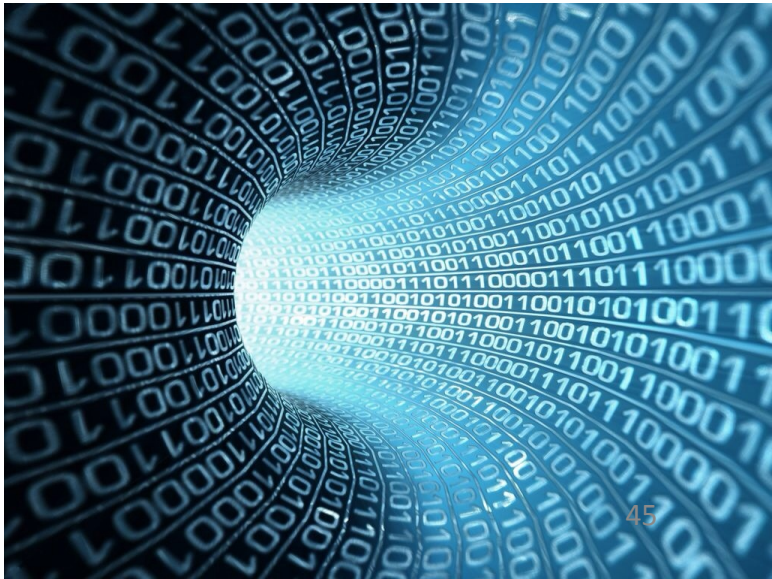


Image from Soumitri Varadarajan blog. Iceberg image © Ralph A. Clevenger. Flickr photo

Scholars' concerns about data

- Research data
  - Funding agency requirements for open data
  - University policies for data ownership
  - Resources and expertise in data-intensive methods
  - Resources for curation and stewardship
  - Open records laws
- Publication data
  - Library management
  - Academic personnel records
  - Evaluation and credit
  - Public-private partnerships



idre UCLA Institute for Digital Research and Education

Search this v

ABOUT PEOPLE PROGRAMS RESOURCES RESEARCH NEWS

Home / Featured / Kent Wada and Christine Borgman Lead Data Governance Task Force

KENT WADA AND CHRISTINE BORGMAN LEAD DATA GOVERNANCE TASK FORCE

February 10, 2015 by Stefanie Pietkiewicz

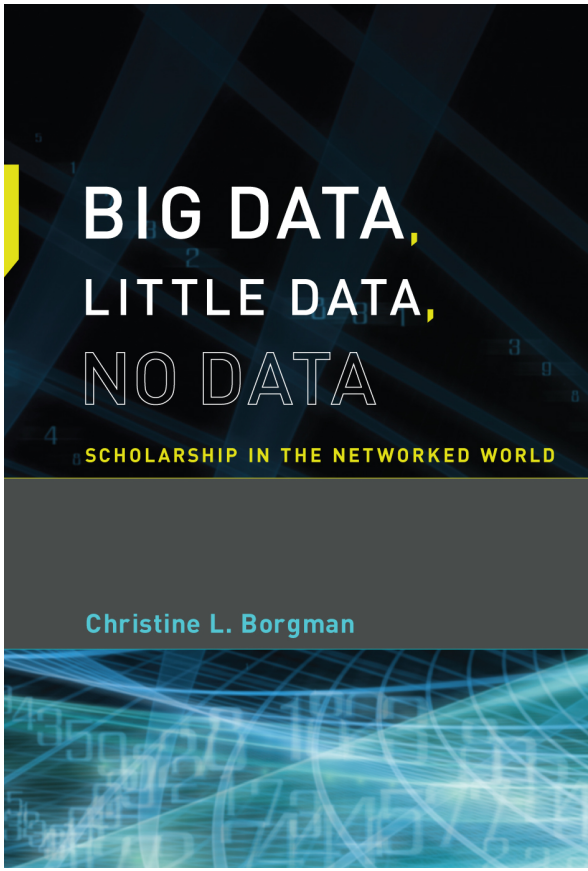
Kent Wada and Christine Borgman

- How should UCLA collect, organize, and use research analytics about our community?
- Who should have access to these data?
  - Within UCLA?
  - In partnership with public and private entities?
- What are the governance principles?
- What are the governance processes?

http://evc.ucla.edu/reports/DGTF-report.pdf

Big Data, Little Data, No Data: Scholarship in the Networked World

- Part I: Data and Scholarship
  - Ch 1: Provocations
  - Ch 2: What Are Data?
  - Ch 3: Data Scholarship
  - Ch 4: Data Diversity
- Part II: Case Studies in Data Scholarship
  - Ch 5: Data Scholarship in the Sciences
  - Ch 6: Data Scholarship in the Social Sciences
  - Ch 7: Data Scholarship in the Humanities
- Part III: Data Policy and Practice
  - Ch 8: Releasing, Sharing, and Reusing Data
  - Ch 9: Credit, Attribution, and Discovery
  - Ch 10: What to Keep and Why



Sustaining Access to Scholarship

depends upon building better

Knowledge Infrastructures

http://knowledgeinfrastructures.org

Acknowledgements

Christine Borgman

Peter Darch

Ashley Sands

Irene Pasquetto

Bernie Randles

Milena Golshan

UCLA Center for Knowledge Infrastructures

NSF

Alfred P. Sloan Foundation 1934